#5

Attorney Docket No. 56055-0013

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

W. Reed HASTINGS, et al.

Serial No.: 09/884,816

Filed on: June 18, 2001

For: APPROACH FOR ESTIMATING USER RATINGS OF ITEMS

OIPE JC134 JAN 03 2002 PATENT & TRADEMARK OFFICE

Group Art Unit No.: 2165

Examiner: Not Yet Assigned

RECEIVED JAN 10 2002 Technology Center 2100

PRELIMINARY AMENDMENT

Commissioner for Patents
Washington, D.C. 20231

Sir:

Prior to examination of the application referenced above, please amend the application referenced above as indicated hereinafter. In accordance with revised 37 C.F.R. § 1.121, a marked up version of the replacement section is provided on separate pages attached to this amendment.

IN THE SPECIFICATION:

Please replace the Brief Description of the Drawings with the following:

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments of the invention are illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements and in which:

FIG. 1 is a diagram depicting an approach for renting items to customers according to an embodiment.

01/07/2002 FFANAEIA 00000051 09884816
01 FC:216 200.00 OP

FIG. 2 is a flow diagram depicting an approach for renting items to customers according to an embodiment.

FIG. 3 is a flow diagram depicting a "Max Out" approach for renting items to customers according to an embodiment.

FIG. 4 is a flow diagram depicting a "Max Turns" approach for renting items to customers according to an embodiment.

FIG. 5 is a diagram depicting an approach for renting audio/video items to customers over the Internet according to an embodiment.

FIG. 6 is a flow diagram illustrating an approach for renting audio/video items to customers over the Internet using both "Max Out" and "Max Turns" according to an embodiment;

FIG. 7 is a block diagram depicting an item space according to an embodiment of the invention;

FIG. 8 is a flow diagram that depicts an approach for estimating how a user would rate an item that the user has not yet rated according to an embodiment of the invention;

FIG. 9 is a block diagram of a computer system upon which embodiments of the invention may be implemented; and

FIG. 10 depicts a possible configuration of components for an internet-based recommendation system;

FIG. 11 is a block diagram that depicts a row of star icons ("stars") used to represent ratings for a movie;

FIG. 12 is a block diagram that depicts a display of no rating and invite users to click on it to input a rating;

FIG. 13 is a block diagram that depicts a display of an output prediction, and invite users to make a selection to refine a recommendations engine database;

FIG. 14 is a block diagram that depicts how a row of stars serves as an input mechanism where a user can click a star to enter a rating;

FIG. 15 is a block diagram that depicts how dynamic image changes may be used to attract user input;

FIG. 16 is a block diagram that depicts a "smooth updates" approach for obtaining user ratings of items;

FIG. 17 is a block diagram that depicts the use of an HTML image table to display star icons to users; and

FIG. 18 is a block diagram that depicts a general model for recommending movie titles using cascading filters and prioritizers.

IN THE SPECIFICATION:

Please replace pages 38 to 56 with the following:

Common User Identities across Different Websites

According to one embodiment of the invention, users may provide a screen-name for their identity that is accessible from all the different web sites using the system. Thus a user who has provided ratings at one Website, may receive predictions at another Website based upon the same set of ratings already provided.

Collaborative Filtering

Novel aspects of the collaborative filtering approach include, without limitation:

- According to one embodiment of the invention, an approach is provided to subdivide the system enabling remote subsystems to share a single database of ratings while retaining high performance;
- According to one embodiment of the invention, a novel user-interface implementation is provided to facilitate collection of ratings from the user, and simultaneously display

predictions to the user for the same item; and

- According to another embodiment of the invention, an approach is provided selecting different items to display based upon inventory and other factors, in conjunction with the predicted ratings of the items.

Multi-Website Implementation

FIG. 10 depicts a possible configuration of components for an internet-based recommendation system 1000. Recommendation System 1000 includes a home site 1002, site 1004 and site 1006. For purposes of explanation, site 1004 is also referred to herein as the "local site" and site 1006 is referred to as the "remote site." Home site 102 includes a user 1008 who interacts with a computer 1010 executing a generic Web browser. Site 1004 includes Web servers 1012, a matching engine 1014, a matching engine 1016, an interface 1018, a database 1020 and an analyzer 1022. Site 1006 includes Web servers 1024, a remote matching engine 1026 and a remote matching engine 1028.

Division of the Computation

The approach for predicting how much a target user will enjoy a particular item includes (a) determining how similar the target user's tastes are to each other user in the database and computing a similarity weight; (b) choosing a set of representative neighboring users who are similar to the target user and who span enough items to make useful predictions; and (c) multiplying how the neighboring users liked the item to be predicted by the similarity weights for the target user, summing and normalizing to get a prediction.

Steps (a) and (b) can be performed offline, since the results of those computations are manageable in size, and relatively stable over time (at least over a period of days to weeks). Step (c) is ideally performed at the time a prediction is required, since the number of possible

target customers and predictable movies is far too large to calculate in advance.

According to one embodiment, calculations (a) and (b) are performed close to the single master database, for example, in analyzer 1022 at site 1004. Remote site 1006 is connected to the site 1004 by a link whose bandwidth is sufficient to download the pre-computed set of neighbors daily, although with high latency.

New user creates account at site1

When a new user visits site 1004 and creates a new account, the new user is invited to offer ratings on items he knows. The ratings are collected by Webservers 1012. Load balancing shares the load between the matching engines, for example, sending customers with odd numbered customer IDs to matching engine 1014, and even numbered customer IDs to matching engine 1016. Matching engines 1014, 1016 record the new customers ratings in database 1020, and initiate computation of the correlation match between the new customer and a pool of pre-selected representative customers who's profiles are loaded into matching engines 1014, 1016. These matches are used to make immediate predictions about what else the new customer might enjoy.

Offline Computation

Later, offline, analyzer 1022 has the opportunity to re-examine the new customer's ratings when updating the pool of representative neighbors. If the new customer has additional information to contribute to the prediction capability of the system 1000, the new customer may become part of the pool of representative neighbors used for making predictions for all other users. Also offline, analyzer 1022 can periodically recomputed the similarity between each user (including the new user) and all the users currently in the representative neighborhood pool. These weightings change gradually as customers rate more movies, and as the members of the

representative pool rate more movies. The weightings for each customer are stored in database 1020.

User returns to site1

When a user returns to site 1004, his rating record and similarity weights are immediately fetched from database 1020. Predictions can be made by applying calculation (c) between the target customer and the pre-loaded pool of representative neighbors.

User returns to site2

If the same customer now visits site 1006 for the first time, he is prompted to offer his screen name or other identifying characteristic, which enables his rating record and similarity weights to be fetched from database 1020. The connection between site 1004 and site 1006 is fast enough to fetch this small record once per customer session with low latency. Site 1006 has already pre-loaded the same set of representative neighbors, and computation proceeds exactly as in the previous case.

User adds more ratings at site2

If the same customer chooses to volunteer more ratings at site 1006, to improve the profile available from which recommendations are made, those ratings will be used locally to refine the weights between the user and the locally stored representative neighbors. Those ratings will also be batched up and later (within minutes or hours) sent back to database 1020, where the offline analyzer 1022 can examine the new data, and determine how the new data affects the selection of representative neighbors for the future. If the new data causes the customer now to be selected as a representative neighbor, then on a subsequent (daily) update of the representative neighborhood pool, his data will be uploaded to the matching engines 1014, 1016 and remote matching engines 1026, 1028.

New user creates account at site2

Similarly, if another new customer creates his account first at site 1006, the local process will initially match that customer up against the preloaded representative neighbors for making predictions immediately, and will later upload the new customer information and ratings to database 1020 for analysis and integration into the overall matching database.

Alternative approaches

Duplicate databases

One alternative approach is to provide a duplicate database at the 1006, so that sites 1004, 1006 can operate independently. A periodic (nightly) process would run to synchronize the data between the two databases. Such a synchronization process would be complex, and would likely require huge bandwidth between the two databases.

Site 2 has remote access to matching engines on site 1

Another alternative places the matching engines dedicated to site 1006 physically at site 1004, close to the single master database 1020 and analyzer 1022. This is computationally and logically simpler, but each rating submission and prediction request requires a round-trip communication between the webservers 1024 at site 1006 and matching engines 1014, 1016 at site 1004. Typical site operations require thousands of predictions a second, requiring huge bandwidth, at very low latency, to make this function with adequate performance (predictions in sub-second times).

Advantages of the proposed approach

Among the benefit of the system 1000 described is that the volume of data to be exchanged with low latency is quite small, and only happens once per customer visit. For

example, a site might have 100,000 visitors per day, each requiring 100 ratings and 100 weights to be uploaded immediately, and resulting in an additional 10 ratings to be sent back later. Each visitor might visit 20 pages, requiring a total of 1000 predictions per visitor, or 100M predictions per day.

Example Embodiments:

4.9 A system to make predictions on user-taste (e.g., customer-taste) at a number of different network resources, e.g., Websites, whose servers are separate, e.g., physically remote from each other.

0 where a master database services recommendation engines at each remote site.

0 where an analysis process at the main site pre-computes a set of representative neighbors for all the sites to use.

0 where a local analysis process at the remote sites performs immediate local matching for prediction for a new customer, but forwards the new customer ratings to the main site for detailed analysis.

0 where the remote sites are connected to the main site via a dedicated data link.

0 where the data link has low latency, but bandwidth inadequate for carrying traffic for every prediction made, and inadequate for replicating and merging the entire database regularly.

0 where a customer's ratings input at any site are used to make predictions for that customer when he visits other sites.

0 where customer's ratings are migrated from one site to another.

0 where a customer's ratings are remotely accessed by another site.

0, 0, 0, 0, and 0 where the items being rated and enjoyment predicted are movies.

0, 0, 0, 0, and 0 where the items being rated and enjoyment predicted are games.

Multi Purpose Graphical User Interface Control

According to one embodiment of the invention, as depicted in FIG. 11, a row of star icons ("stars") is used to represent ratings for a movie. There are several dimensions that may be displayed as stars:

- A user's input rating on a specific movie
- A personal, individualized prediction about a specific movie
- The average rating of all users on a specific movie
- Editorial ratings (such as by well-known critics) on a specific movie.

When a particular user is unknown, a personalized prediction or rating cannot be displayed. Once a user has rated sufficient other movies to provide confidence in predicting a rating for this movie, a prediction can be shown. If the user has seen the title, our prediction may differ from the user's actual level of enjoyment, and the user can enter his actual rating to help train the recommendation engine for future predictions. If the user has entered a rating, the prediction will be identical to his rating.

According to one embodiment of the invention, a graphical interface device is used to:

- As depicted in FIG. 12, display no rating and invite users to click on it to input a rating:
- As depicted in FIG. 13, display an output prediction, and still invite users to select a rating to help refine a recommendations engine database:

As depicted in FIG. 14, a row of stars serves as an input mechanism where a user can click a star to enter a rating:

A single row of stars may be used that are empty outlines for the no-ratings case, are colored red for an output prediction, and colored gold to represent an input rating.

Dynamic image changes to attract input

One feature is making the image obviously a way to input a rating too. This may be accomplished by dynamically changing the image as the user's mouse passes over the image. As mouse moves over a particular star, the predicted rating red filled stars disappear, and the star under the mouse and all stars to the left of it are replaced with the gold input rating star. Simultaneously, the words above the image change to "click rate movie". As the mouse drags left and right over the stars, the appropriate star outlines are outlined or filled in to give the impression that the user is dynamically changing the rating. FIG. 15 is a block diagram that depicts how dynamic image changes may be used to attract user input.

Smooth updates

According to the "smooth updates" approach, when the user finally clicks on a star, the JavaScript redisplays the stars image displaying the users input, and revises the adjacent words to show an input accepted. Also, the web browser sends the input rating to the web server (by posting a form) and waits for a response. Normally, this would cause the web server to compute a new page and send it to the browser for display, which involves a round-trip communication which could be anywhere from fractions of a second to tens of seconds if the user is on a slower dial-up line.

One aspect of this approach is for users to be able to enter many ratings as smoothly as possible, so a novel technique is used to not update the web browser window. The web server first records the user's input. The web protocols require that it respond, which it does by creating a new page and requiring that the client machine create a new window to display that page. The new window is as small as possible, and where possible is positioned off the visible screen of the client machine. The page that is sent to the new window contains only the

JavaScript commands to make the window invisible. Together, these techniques ensure that the end-user is only minimally aware that anything was repainted or redisplayed.

Display of confidence

According to another embodiment of the invention, the prediction engine provides a confidence level for each individualized predictions. Confidence information may be separately represented using color to represent confidence (from watery pastels for low confidence to bright primary colors for high confidence). An alternative implementation is to use size of the images to represent different confidence levels, or a different graphical scale alongside.

Implementation Examples

According to one embodiment, as depicted in FIG. 16, the rows of stars 1602, 1604, 1606 are assembled as an HTML table 1600 with 10 slots. JavaScript selects appropriate images to fill the cells of the table, and responds to mouse movement by dynamically changing the images to present the proper animation and pictures. There are six images, being the half stars each in white, gold, and red.

In an alternative embodiment, as depicted in a table 1700 of FIG. 17, there are 21 distinct images for the whole row of stars, representing the 5 empty stars image, and ten versions each of gold and red stars. An Image Map maps regions of the single bar of stars image into regions, and JavaScript selects and loads the appropriate composite image based upon which region the mouse is over.

According to another embodiment of the invention, star images are used that have transparent centers, where the color changes are effected by the JavaScript by varying the background color of the table cells where the stars are laid out. In the figures above, the 5 stars permit 10 or 11 possible values by permitting half-stars to be colored. The implementation

could equally well use whole stars (5 or 6 ratings) or other fractions for different granularities.

Example Embodiments:

5.6.1 Use of a single graphical user interface device to display output and capture input on a web-browser or other type of transaction-oriented system.

0 where the output display (e.g. a prediction) is of a different parameter than the input parameter (e.g. of a rating).

0 where a linear graphical display is used to input quantized numerical ratings.

0, 0, and 0 where confidence in results is displayed graphically

0 where color and/or size is used for confidence display

Technique for simulating one-way updates of data from the web-browser to the web-server without requiring a page update on the web-browser.

0 based on the web-server sending a minimal response to a separate browser window hidden from the user's view.

0 where the response is sent to an embedded iframe in the browser

0 where the iframe displays a relevant response, such as a count of the number of customer inputs, that changes as the customer interacts with the interface.

0 using client-side JavaScript to update the page seen by the user immediately, avoiding the need for a round-trip delay to the web-server.

0 - 0 where the user interface is connected to a system for accepting ratings for and making predictions on taste for items.

0 where the items being rated and predicted are movies.

0 where the items being rated and predicted are games.

Techniques for Inventory Management for an Electronic Commerce Rental Business

An on-line electronic-commerce business has exceptional ability dynamically to tailor its merchandising (or more generally, its product presentation, advertising, and promotional displays) to particular individuals who visit the store web site. This invention is a process for managing different presentations of product for each visitor to a store with a view to managing inventory effectively.

Consider the example of a movie rental business. Customers choose DVD movies at an electronic-commerce web site. Movies are stored at a central warehouse, and are shipped by US mail anywhere in the US. A key business concern is satisfying the biggest possible demand for a specific title with the smallest possible number of copies of that title.

As in a physical store, product positioned in certain strategic locations in the store is consumed more frequently than if the same product is positioned in less visible parts of the store. By modulating which titles get the key positions in the store, the demand for each title can be controlled dramatically. In an electronic commerce store, the key locations are the home page and the other top-level "face pages" that feature a small number of movies along with box shots and editorial commentary.

For each face page (more generally, for each context), the titles to be displayed are changed for

each customer who visits the store, and each visit they make, based upon a variety of factors. The web-server is programmed to consider all these factors, select appropriate movies, construct a page dynamically, and send the HTML code for that page to the customer's web browser for display. Each time the page is requested, a different result can be generated.

According to one embodiment of the invention, a general model 1800 for selection of titles is logically viewed as a series of cascading filters and prioritizers 1804-1812:

Consider a section of a face page presenting a feature about movies with great soundtracks. Starting from the catalog 1802 of all DVD movies available, a filter 1804 picks the movies chosen by the store editorial staff as being the movies with great sound. This is done by associating a context keyword ("great_sound") against the catalog entry in the movie content database. Filter 1804 picks out all titles from catalog 1802 with that associated keyword. The next filter 1806 eliminates (or equivalently, strongly downgrades in priority) all titles where there is not a minimum level of stock. If there are zero copies in inventory, there is no point in using valuable store space to present it to a potential customer.

The third filter 1808 eliminates titles known to have been previously rented or vie by the particular customer. For a new visitor with no known history, this filter 1808 is null course. Customers wishing to rent a title a second time can still find it by searching, or by reviewing their previous rental history. Priority space in the store should be focused on tit likely to be most appealing to each individual visitor. If a customer has already ordered or queued up particular titles for future delivery, these too can be excluded at this stage.

The fourth step is the recommendation engine 1810. This matches the known ratings of the current visitor against a large database of other visitors, and ranks titles by expected appeal to the target visitor, taking into account the confidence of the predictions from the engine. This

step does not eliminate any titles, it simply brings to the top of the list the titles expected to be most appealing. For a new visitor with no ratings in the database against which to match his tastes, this filter simply ranks the titles by average ratings across the whole population of customers who have rated each title in the filter list.

The fifth filter 1812 step examines days' supply outstanding for each title: this is the number of copies on hand, divided by the average of the daily consumption of the title during the previous days, considering the most recent 14 days of shipping behavior. This figure approximates the number of days that the current supply would last, assuming no change in recommendation priority, and no returned copies. For the titles that enter this filter step with all other factors being equal, those titles that have more days outstanding of stock will be elevated in priority over those having fewer. The reason for using DSO as the priority metric instead of just current inventory level is so that equal consideration is given to a specialty interest title with small total inventory and small average run-rate as is given to a broad-appeal title with huge inventory and huge average run-rate. Other information about pending shipments from the warehouse, such as orders for items for future delivery, can also be applied to the DSO metric. What matters is that there be a prioritization scheme that measures the ratio of inventory to some metric of actual or predicted run-rate which can be used to modulate presentation of otherwise attractive items to the customer.

This last filter 1812 step also acts as a governor on new-release promotion. For the first few weeks of a new title's release, natural demand, un-augmented by featured presentations of the title on face pages, drives high average run rates. At the same time, the number of copies on hand is currently low, as the proper number is purchased incrementally to respond to the actual demand. As the newness of the title fades, the demand from customers actively seeking out the

title fades, and the DSO number climbs rapidly, elevating the priority of the title. This happens with similar priority regardless of whether the title has been a 1,000-a-day title accumulating 20,000 copies in inventory, or a 10-a-day title accumulating 2,000 copies in inventory.

For customer trust, it is important that the inventory prioritizer have much lower influence on the titles presented than the recommendation engine predicting titles for the specific customer based upon his recorded ratings of other movies. A logarithmic transfer function is also appropriate on this filter too, as differences near the low-end of DSO are far more significant than differences at the higher DSO numbers.

Alternative Embodiments

With this approach, a relatively large amount of demand may be distributed over a small number of copies of a particular title, which is appropriate for a rental business where the cost of a rental is minimized by increasing the number of turns and the length of time for which it turns. For a product sell-through application, the filters 1804-1812 might emphasize other aspects of inventory, such as the proximity of volume purchase discounts, opportunities to feature titles by a distributor with whom especially favorable terms have been negotiated, or even accessibility of particular titles in a complex network of warehouses (emphasizing titles in the short-term or high-cost zones of the warehouse).

This approach is applicable wherever product can be differently featured to individual consumers, or small classes of consumers. The example application described above uses an electronic commerce Website to present different rental items to each customer. An alternative application domain related to direct marketing: one-to-one outbound advertising

carried by electronic mail, or physical mail, where each piece is tailored to the specific recipient based upon his tastes and available inventory.

REMARKS

By this amendment, pages 38 to 56 have been amended to remove figures from the specification and to add references to the figures that are now provided herewith on separate sheets. The "Brief Description of the Drawings" section was also amended to reference the figures provided on separate sheets. It is respectfully submitted that these amendments do not add any new matter to this application.

The Examiner is invited to contact the undersigned by telephone if it is believed that such contact would further the examination of the present application.

If there are any additional charges, please charge them to our Deposit Account No. 50-1302.

Respectfully submitted,

HICKMAN PALERMO TRUONG & BECKER LLP

Dated: November 9, 2001

Edward A. Becker
Reg. No. 37,777

1600 Willow Street
San Jose, California 95125-5106
Telephone: (408) 414-1204
Facsimile: (408) 414-1076

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Commissioner for Patents, Washington, DC 20231

on November 9, 2001 by Sheila Severinghaus

Sheila Severinghaus

OIPE JC134 JAN 0 3 2002 PATENT & TRADEMARK OFFICE

MARKED UP VERSION OF REPLACEMENT SECTIONS

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments of the invention are illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements and in which:

FIG. 1 is a diagram depicting an approach for renting items to customers according to an embodiment.

FIG. 2 is a flow diagram depicting an approach for renting items to customers according to an embodiment.

FIG. 3 is a flow diagram depicting a "Max Out" approach for renting items to customers according to an embodiment.

FIG. 4 is a flow diagram depicting a "Max Turns" approach for renting items to customers according to an embodiment.

FIG. 5 is a diagram depicting an approach for renting audio/video items to customers over the Internet according to an embodiment.

FIG. 6 is a flow diagram illustrating an approach for renting audio/video items to customers over the Internet using both "Max Out" and "Max Turns" according to an embodiment;

FIG. 7 is a block diagram depicting an item space according to an embodiment of the invention;

FIG. 8 is a flow diagram that depicts an approach for estimating how a user would rate an item that the user has not yet rated according to an embodiment of the invention; ~~and~~

FIG. 9 is a block diagram of a computer system upon which embodiments of the invention may be implemented; and

FIG. 10 depicts a possible configuration of components for an internet-based recommendation system;

FIG. 11 is a block diagram that depicts a row of star icons ("stars") used to represent ratings for a movie;

FIG. 12 is a block diagram that depicts a display of no rating and invite users to click on it to input a rating;

FIG. 13 is a block diagram that depicts a display of an output prediction, and invite users to make a selection to refine a recommendations engine database;

FIG. 14 is a block diagram that depicts how a row of stars serves as an input mechanism where a user can click a star to enter a rating;

FIG. 15 is a block diagram that depicts how dynamic image changes may be used to attract user input;

FIG. 16 is a block diagram that depicts a "smooth updates" approach for obtaining user ratings of items;

FIG. 17 is a block diagram that depicts the use of an HTML image table to display star icons to users; and

FIG. 18 is a block diagram that depicts a general model for recommending movie titles using cascading filters and prioritizers.

Common User Identities across Different Websites

According to one embodiment of the invention, users may provide a screen-name for their identity that is accessible from all the different web sites using the system. Thus a user who has provided ratings at one Website, may receive predictions at another Website based upon the same set of ratings already provided.

Collaborative Filtering

Novel aspects of the collaborative filtering approach include, without limitation:

- According to one embodiment of the invention, an approach is provided to subdivide the system enabling remote subsystems to share a single database of ratings while retaining high performance;
- According to one embodiment of the invention, a novel user-interface implementation is provided to facilitate collection of ratings from the user, and simultaneously display predictions to the user for the same item; and
- According to another embodiment of the invention, an approach is provided selecting different items to display based upon inventory and other factors, in conjunction with the predicted ratings of the items.

Multi-Website Implementation

~~The diagram below shows~~ FIG. 10 depicts a possible configuration of components for an internet-based recommendation system 1000. Recommendation System 1000 includes a home site 1002, site 1004 and site 1006. For purposes of explanation, site 1004 is also referred to herein as the "local site" and site 1006 is referred to as the "remote site." Home site 102 includes a user 1008 who interacts with a computer 1010 executing a generic Web browser.

Site 1004 includes Web servers 1012, a matching engine 1014, a matching engine 1016, an interface 1018, a database 1020 and an analyzer 1022. Site 1006 includes Web servers 1024, a remote matching engine 1026 and a remote matching engine 1028.

Division of the Computation

The approach for predicting how much a target user will enjoy a particular item includes (a) determining how similar the target user's tastes are to each other user in the database and computing a similarity weight; (b) choosing a set of representative neighboring users who are similar to the target user and who span enough items to make useful predictions; and (c) multiplying how the neighboring users liked the item to be predicted by the similarity weights for the target user, summing and normalizing to get a prediction.

Steps (a) and (b) can be performed offline, since the results of those computations are manageable in size, and relatively stable over time (at least over a period of days to weeks). Step (c) is ideally performed at the time a prediction is required, since the number of possible target customers and predictable movies is far too large to calculate in advance.

According to one embodiment, calculations (a) and (b) are performed close to the single master database, for example, in ~~the~~ analyzer 1022 ~~component~~ at ~~the main~~ site 1004. ~~The~~ ~~r~~Remote site 1006~~2~~ is connected to the ~~main~~ site 1004 by a link whose bandwidth is sufficient to download the pre-computed set of neighbors daily, although with high latency.

New user creates account at site ~~1~~1

When a new user visits site 1004 and creates a new account, the new user is invited to offer ratings on items he knows. The ratings are collected by ~~the~~ Webservers 1012. Load balancing shares the load between the matching engines, for example, sending customers with odd numbered customer IDs to matching engine 1014, and even numbered customer IDs to

matching engine ~~2~~1016. ~~The m~~Matching engines 1014, 1016 records the new customers ratings in ~~the~~ database 1020, and initiates computation of the correlation match between the new customer and a pool of pre-selected representative customers who's profiles are loaded into ~~the~~ matching engines 1014, 1016. These matches are used to make immediate predictions about what else the new customer might enjoy.

Offline Computation

Later, offline, ~~the~~ analyzer 1022 has the opportunity to re-examine the new customer's ratings when updating the pool of representative neighbors. If the new customer has additional information to contribute to the prediction capability of the system 1000, the new customer may become part of the pool of representative neighbors used for making predictions for all other users. Also offline, ~~the~~ analyzer 1022 can periodically recomputed the similarity between each user (including the new user) and all the users currently in the representative neighborhood pool. These weightings change gradually as customers rate more movies, and as the members of the representative pool rate more movies. The weightings for each customer are stored in ~~the~~ database 1020.

User returns to site ~~1~~1

When a user returns to site 1004, his rating record and similarity weights are immediately fetched from ~~the~~ database 1020. Predictions can be made by applying calculation (c) between the target customer and the pre-loaded pool of representative neighbors.

User returns to site ~~2~~2

If the same customer now visits site ~~2~~1006 for the first time, he is prompted to offer his screen name or other identifying characteristic, which enables his rating record and similarity weights to be fetched from ~~the~~ database 1020~~at site 1~~. The connection between site 1004 and

site 10062 is fast enough to fetch this small record once per customer session with low latency. Site 10062 has already pre-loaded the same set of representative neighbors, and computation proceeds exactly as in the previous case.

User adds more ratings at site 22

If the same customer chooses to volunteer more ratings at site 10062, to improve the profile available from which recommendations are made, those ratings will be used locally to refine the weights between the user and the locally stored representative neighbors. Those ratings will also be batched up and later (within minutes or hours) sent back to the master database 1020, where the offline analyzer 1022 can examine the new data, and determine how the new data affects the selection of representative neighbors for the future. If the new data causes the customer now to be selected as a representative neighbor, then on a subsequent (daily) update of the representative neighborhood pool, his data will be uploaded to the local and matching engines 1014, 1016 and remote matching engines 1026, 1028.

New user creates account at site 22

Similarly, if another new customer creates his account first at site 10062, the local process will initially match that customer up against the preloaded representative neighbors for making predictions immediately, and will later upload the new customer information and ratings to the master database 1020 for analysis and integration into the overall matching database.

Alternative approaches

Duplicate databases

One alternative approach is to provide a duplicate database at the remote site 21006, so that sites 1004, 1006 1 and site 2 could can operate independently. A periodic (nightly) process

would run to synchronize the data between the two databases. Such a synchronization process would be complex, and would likely require huge bandwidth between the two databases.

Site 2 has remote access to matching engines on site 1

Another alternative places the matching engines dedicated to site 10062 physically at site 1004, close to the single master database 1020 and analyzer 1022~~sis process~~. This is computationally and logically simpler, but each rating submission and prediction request requires a round-trip communication between the webservers 1024 at site 10062 and ~~the~~ matching engines 1014, 1016 at site 1004. Typical site operations require thousands of predictions a second, requiring huge bandwidth, at very low latency, to make this function with adequate performance (predictions in sub-second times).

Advantages of the proposed approach

Among the benefit of the ~~architecture~~ system 1000 described is that the volume of data to be exchanged with low latency is quite small, and only happens once per customer visit. For example, a site might have 100,000 visitors per day, each requiring 100 ratings and 100 weights to be uploaded immediately, and resulting in an additional 10 ratings to be sent back later. Each visitor might visit 20 pages, requiring a total of 1000 predictions per visitor, or 100M predictions per day.

Example Embodiments:

4.9 A system to make predictions on user-taste (e.g., customer-taste) at a number of different network resources, e.g., Websites, whose servers are separate, e.g., physically remote from each other.

0 where a master database services recommendation engines at each remote site.

0 where an analysis process at the main site pre-computes a set of representative neighbors for all the sites to use.

0 where a local analysis process at the remote sites performs immediate local matching for prediction for a new customer, but forwards the new customer ratings to the main site for detailed analysis.

0 where the remote sites are connected to the main site via a dedicated data link.

0 where the data link has low latency, but bandwidth inadequate for carrying traffic for every prediction made, and inadequate for replicating and merging the entire database regularly.

0 where a customer's ratings input at any site are used to make predictions for that customer when he visits other sites.

0 where customer's ratings are migrated from one site to another.

0 where a customer's ratings are remotely accessed by another site.

0, 0, 0, 0, and 0 where the items being rated and enjoyment predicted are movies.

0, 0, 0, 0, and 0 where the items being rated and enjoyment predicted are games.

Multi Purpose Graphical User Interface Control

According to one embodiment of the invention, as depicted in FIG. 11, a row of star icons ("stars")s is used to represent ratings for a movie. There are several dimensions that may be displayed as stars:

- A user's input rating on a specific movie
- A personal, individualized prediction about a specific movie
- The average rating of all users on a specific movie
- Editorial ratings (such as by well-known critics) on a specific movie.

When a particular user is unknown, a personalized prediction or rating cannot be displayed. Once a user has rated sufficient other movies to provide confidence in predicting a rating for this movie, a prediction can be shown. If the user has seen the title, our prediction may differ from the user's actual level of enjoyment, and the user can enter his actual rating to help train the recommendation engine for future predictions. If the user has entered a rating, the prediction will be identical to his rating.

According to one embodiment of the invention, a graphical interface device is used to:

- As depicted in FIG. 12, ~~D~~display no rating and invite users to click on it to input a rating:
- As depicted in FIG. 13, ~~D~~display an output prediction, and still invite users to select a rating ~~click on it~~ to help refine a~~our~~ recommendations engine database:

As depicted in FIG. 14, a row of stars s~~S~~erves as an input mechanism ~~device~~ where a user can click a star to enter a rating:

A single row of stars may be used that are empty outlines for the no-ratings case, are colored red for an output prediction, and colored gold to represent an input rating.

Dynamic image changes to attract input

One feature is making the image obviously a way to input a rating too. This may be accomplished by dynamically changing the image as the user's mouse passes over the image. As mouse moves over a particular star, the predicted rating red filled stars disappear, and the star under the mouse and all stars to the left of it are replaced with the gold input rating star. Simultaneously, the words above the image change to "click rate movie". As the mouse drags left and right over the stars, the appropriate star outlines are outlined or filled in to give the impression that the user is dynamically changing the rating. FIG. 15 is a block diagram that depicts how dynamic image changes may be used to attract user input.

Smooth updates

According to the "smooth updates" approach, when the user finally clicks on a star, the JavaScript redisplays the stars image displaying the users input, and revises the adjacent words to show an input accepted. Also, the web browser sends the input rating to the web server (by posting a form) and waits for a response. Normally, this would cause the web server to compute a new page and send it to the browser for display, which involves a round-trip communication which could be anywhere from fractions of a second to tens of seconds if the user is on a slower dial-up line.

One aspect of this approach is for users to be able to enter many ratings as smoothly as possible, so a novel technique is used to not update the web browser window. The web server first records the user's input. The web protocols require that it respond, which it does by creating a new page and requiring that the client machine create a new window to display that

page. The new window is as small as possible, and where possible is positioned off the visible screen of the client machine. The page that is sent to the new window contains only the JavaScript commands to make the window invisible. Together, these techniques ensure that the end-user is only minimally aware that anything was repainted or redisplayed.

Display of confidence

According to another embodiment of the invention, the prediction engine provides a confidence level for each individualized predictions. Confidence information may be separately represented using color to represent confidence (from watery pastels for low confidence to bright primary colors for high confidence). An alternative implementation is to use size of the images to represent different confidence levels, or a different graphical scale alongside.

Implementation Examples

According to one embodiment, as depicted in FIG. 16, the rows of stars 1602, 1604, 1606 are assembled as an HTML table 1600 with 10 slots. JavaScript selects appropriate images to fill the cells of the table, and responds to mouse movement by dynamically changing the images to present the proper animation and pictures. There are six images, being the half stars each in white, gold, and red.

In an alternative embodiment, as depicted in a table 1700 of FIG. 17, there are 21 distinct images for the whole row of stars, representing the 5 empty stars image, and ten versions each of gold and red stars. An Image Map maps regions of the single bar of stars image into regions, and JavaScript selects and loads the appropriate composite image based upon which region the mouse is over.

According to another embodiment of the invention, star images are used that have transparent centers, where the color changes are effected by the JavaScript by varying the

background color of the table cells where the stars are laid out. In the figures above, the 5 stars permit 10 or 11 possible values by permitting half-stars to be colored. The implementation could equally well use whole stars (5 or 6 ratings) or other fractions for different granularities.

Example Embodiments:

5.6.1 Use of a single graphical user interface device to display output and capture input on a web-browser or other type of transaction-oriented system.

0 where the output display (e.g. a prediction) is of a different parameter than the input parameter (e.g. of a rating).

0 where a linear graphical display is used to input quantized numerical ratings.

0, 0, and 0 where confidence in results is displayed graphically

0 where color and/or size is used for confidence display

Technique for simulating one-way updates of data from the web-browser to the web-server without requiring a page update on the web-browser.

0 based on the web-server sending a minimal response to a separate browser window hidden from the user's view.

0 where the response is sent to an embedded iframe in the browser

0 where the iframe displays a relevant response, such as a count of the number of customer inputs, that changes as the customer interacts with the interface.

0 using client-side JavaScript to update the page seen by the user immediately, avoiding the need for a round-trip delay to the web-server.

0 - 0 where the user interface is connected to a system for accepting ratings for and making predictions on taste for items.

0 where the items being rated and predicted are movies.

0 where the items being rated and predicted are games.

Techniques for Inventory Management for an Electronic Commerce Rental Business

An on-line electronic-commerce business has exceptional ability dynamically to tailor its merchandising (or more generally, its product presentation, advertising, and promotional displays) to particular individuals who visit the store web site. This invention is a process for managing different presentations of product for each visitor to a store with a view to managing inventory effectively.

Consider the example of a movie rental business. Customers choose DVD movies at an electronic-commerce web site. Movies are stored at a central warehouse, and are shipped by US mail anywhere in the US. A key business concern is satisfying the biggest possible demand for a specific title with the smallest possible number of copies of that title.

As in a physical store, product positioned in certain strategic locations in the store is consumed more frequently than if the same product is positioned in less visible parts of the store. By modulating which titles get the key positions in the store, the demand for each title can be controlled dramatically. In an electronic commerce store, the key locations are the home page and the other top-level "face pages" that feature a small number of movies along with box shots and editorial commentary.

For each face page (more generally, for each context), the titles to be displayed are changed for

each customer who visits the store, and each visit they make, based upon a variety of factors. The web-server is programmed to consider all these factors, select appropriate movies, construct a page dynamically, and send the HTML code for that page to the customer's web browser for display. Each time the page is requested, a different result can be generated.

According to one embodiment of the invention, ~~a~~the general model 1800 for selection of titles is logically viewed as a series of cascading filters and prioritizers 1804-1812:

Consider a section of a face page presenting a feature about movies with great soundtracks. Starting from the catalog 1802 of all DVD movies available, a filter 1804 picks the movies chosen by the store editorial staff as being the movies with great sound. This is done by associating a context keyword ("great_sound") against the catalog entry in the movie content database. ~~The f~~Filter 1804 picks out all titles from ~~the~~ catalog 1802 with that associated keyword.

The next filter 1806 eliminates (or equivalently, strongly downgrades in priority) all titles where there is not a minimum level of stock. If there are zero copies in inventory, there is no point in using valuable store space to present it to a potential customer.

The third filter 1808 eliminates titles known to have been previously rented or viewed by the particular customer. For a new visitor with no known history, this filter 1808 is null, of course. Customers wishing to rent a title a second time can still find it by searching, or by reviewing their previous rental history. Priority space in the store should be focused on titles likely to be most appealing to each individual visitor. If a customer has already ordered or queued up particular titles for future delivery, these too can be excluded at this stage.

The fourth step is the recommendation engine 1810. This matches the known ratings of the current visitor against a large database of other visitors, and ranks titles by expected appeal

to the target visitor, taking into account the confidence of the predictions from the engine. This step does not eliminate any titles, it simply brings to the top of the list the titles expected to be most appealing. For a new visitor with no ratings in the database against which to match his tastes, this filter simply ranks the titles by average ratings across the whole population of customers who have rated each title in the filter list.

The fifth filter 1812 step examines days' supply outstanding for each title: this is the number of copies on hand, divided by the average of the daily consumption of the title during the previous days, considering the most recent 14 days of shipping behavior. This figure approximates the number of days that the current supply would last, assuming no change in recommendation priority, and no returned copies. For the titles that enter this filter step with all other factors being equal, those titles that have more days outstanding of stock will be elevated in priority over those having fewer. The reason for using DSO as the priority metric instead of just current inventory level is so that equal consideration is given to a specialty interest title with small total inventory and small average run-rate as is given to a broad-appeal title with huge inventory and huge average run-rate. Other information about pending shipments from the warehouse, such as orders for items for future delivery, can also be applied to the DSO metric. What matters is that there be a prioritization scheme that measures the ratio of inventory to some metric of actual or predicted run-rate which can be used to modulate presentation of otherwise attractive items to the customer.

This last filter 1812 step also acts as a governor on new-release promotion. For the first few weeks of a new title's release, natural demand, un-augmented by featured presentations of the title on face pages, drives high average run rates. At the same time, the number of copies on hand is currently low, as the proper ~~number are~~number is purchased incrementally to respond to

the actual demand. As the newness of the title fades, the demand from customers actively seeking out the title fades, and the DSO number climbs rapidly, elevating the priority of the title. This happens with similar priority regardless of whether the title has been a 1,000-a-day title accumulating 20,000 copies in inventory, or a 10-a-day title accumulating 2,000 copies in inventory.

For customer trust, it is important that the inventory prioritizer have much lower influence on the titles presented than the recommendation engine predicting titles for the specific customer based upon his recorded ratings of other movies. A logarithmic transfer function is also appropriate on this filter too, as differences near the low-end of DSO are far more significant than differences at the higher DSO numbers.

Alternative Embodiments

With this approach, a relatively large amount of demand may be distributed over a small number of copies of a particular title, which is appropriate for a rental business where the cost of a rental is minimized by increasing the number of turns and the length of time for which it turns. For a product sell-through application, the filters 1804-1812 might emphasize other aspects of inventory, such as the proximity of volume purchase discounts, opportunities to feature titles by a distributor with whom especially favorable terms have been negotiated, or even accessibility of particular titles in a complex network of warehouses (emphasizing titles in the short-term or high-cost zones of the warehouse).

This approach is applicable wherever product can be differently featured to individual consumers, or small classes of consumers. The example application described above uses an electronic commerce Website to present different rental items to each

customer. An alternative application domain related to direct marketing: one-to-one outbound advertising carried by electronic mail, or physical mail, where each piece is tailored to the specific recipient based upon his tastes and available inventory.